The Roots of Performance Disparity in Multilingual Language Models: Intrinsic Modeling Difficulty or Design Choices?

Chen Shani[1], Yuval Reif[2], Nathan Roll[1], Dan Jurafsky[1], Ekaterina Shutova[3]
[1]Stanford University, [2]The Hebrew University of Jerusalem, [3]University of Amsterdam

Abstract

Multilingual language models (LMs) promise broader NLP access, yet current systems deliver uneven performance across the world's languages. This survey examines why these gaps persist and whether they reflect intrinsic linguistic difficulty or modeling artifacts. We organize the literature around two questions: do linguistic disparities arise from representation and allocation choices (e.g., tokenization, encoding, data exposure, parameter sharing) rather than inherent complexity; and which design choices mitigate inequities across typologically diverse languages. We review linguistic features, such as orthography, morphology, lexical diversity, syntax, information density, and typological distance, linking each to concrete modeling mechanisms. Gaps often shrink when segmentation, encoding, and data exposure are normalized, suggesting much apparent difficulty stems from current modeling choices. We synthesize these insights into design recommendations for tokenization, sampling, architectures, and evaluation to support more balanced multilingual LMs.

1 Introduction

Multilingual LMs have expanded NLP's reach by enabling a single model to perform tasks across many languages. They are pretrained on text from hundreds of languages, sharing parameters and representations (Devlin et al., 2019; Conneau et al., 2020a; Le Scao et al., 2022; Imani et al., 2023; Dang et al., 2024). This enables cross-lingual transfer, where patterns learned in one language improve performance in others (Pires et al., 2019; Conneau et al., 2020b; Lauscher et al., 2020; Malkin et al., 2022; Blevins et al., 2024). Despite these advantages, persistent performance disparities across languages limit the practical reach of multilingual models (Wang et al., 2025; Ghosh et al., 2025).

These disparities systematically follow cross-linguistic patterns: higher-resource languages and those structurally similar to dominant training languages generally perform better than low-resource or typologically distant ones (Zhao et al., 2025; Akindotuni, 2025). The disparities often persist even with large-scale pretraining, suggesting that scaling alone cannot ensure equitable performance (Hoffmann et al., 2022; He et al., 2025). This raises a central question: **are some languages inherently harder to model, or do performance gaps reflect engineering artifacts and design choices?**

We review how linguistic structure interacts with multilingual design choices to shape performance gaps via two questions: whether disparities stem from intrinsic difficulty or modeling artifacts (e.g., tokenization, data allocation, shared-parameter interference); and which design choices mitigate inequities. We consolidate our findings into a set of recommendations for tokenization, data sampling, model architectures, and evaluation, highlighting where evaluations confound learnability with tokenization or encoding artifacts (Table 1).

Our synthesis suggests that cross-linguistic gaps rarely reflect intrinsic modeling complexity. Instead, they arise via three mechanisms: (1) shared-parameter training induces negative transfer when typological diversity exceeds effective capacity (Pfeiffer et al., 2022; Blevins et al., 2024; Chang et al., 2024a); (2) tokenization and encoding fragment words or penalize byte-heavy scripts, inflating sequence length without added meaning (Rust et al., 2021; Arnett et al., 2024; Lundin et al., 2025; Land and Arnett, 2025); and (3) data sampling and evaluation misrepresent semantic exposure. Gaps shrink when normalizing segmentation, encoding, and exposure or explicitly allocating capacity, indicating that difficulty stems from modeling choices.

This first systematic review of cross-linguistic modeling difficulty research offers practical design recommendations for multilingual LMs to achieve balanced performance across diverse languages.

Linguistic Factor	Observed Artifact	Modeling Mechanism	Design Levers
Orthography and encoding granularity (§2.1)	Encoding inefficiency (byte premium); reduced effective exposure under fixed budgets; inconsistent written signal	UTF-8 byte-length asymmetries inflate sequence length and reduce effective training signal for many non-Latin scripts	Byte-normalized sampling; script-aware or language-adaptive tokenization; alternative encodings; tokenizer-free models (§3.1, §3.2)
Morphology: productivity and compounding (§2.2, §2.3)	Tokenization (over-segmentation, longer sequences); diluted training signal across surface forms	BPE subwords misalign with morpheme boundaries, yielding inconsistent tokenization	Morphology-aware tokenization; language-aware vocabularies (§3.1)
Information density and redundancy (§2.5)	Unequal semantic coverage under fixed token budgets	Token-based budgeting allocates unequal information per unit of training, confounding cross-language comparisons	Information-, byte- or morpheme-normalized sampling; adaptive scaling of effective exposure to language data (§3.2, §3.4)
Typological and syntactic divergence (§2.6, §2.4)	Negative transfer under shared parameters; degraded syntactic generalization	Shared capacity induces gradient conflict and representation collapse when languages differ strongly in structure	Modular capacity and language adapters; typology-aware routing; controlled sharing (§3.4)
Evaluation sensitivity to tokenization and encoding (§2.1, §2.5)	Perplexity comparability confounded by segmentation and byte length	Subword-level metrics conflate segmentation decisions with predictability	Report character/morpheme-level metrics; tokenization diagnostics and typology-aware probes (§3.3)

Table 1: Linking linguistic properties to multilingual modeling artifacts, mechanisms, and design levers; section references point to the supporting evidence discussed in the survey.

2 Linguistic Properties

Human languages evolve under multiple, sometimes competing objectives, producing systematic trade-offs across morphology, syntax, and phonology (Gibson et al., 2019). Information may be densely packed within words or distributed across syntax; flexible word order can be balanced by overt marking such as case or agreement. These features preserve overall communicative efficiency despite wide typological diversity (Gibson et al., 2019; Lian et al., 2023). Human acquisition aligns with this view: children reliably acquire their ambient language, though the timing and difficulty of specific constructions vary by typology rather than defining a universal hierarchy of "hard" languages (Slobin, 1987; Berman, 2014).

We review linguistic properties associated with cross-linguistic performance variation, where *learnability* refers to sample efficiency and predictive performance (perplexity, downstream accuracy). Drawing on NLP, computational linguistics, typology, and information theory, we show how these properties influence tokenization, data allocation, and architecture in multilingual LMs. Each subsection defines a property, summarizes evidence, and discusses factors affecting modeling success.

2.1 Orthography

Orthography drives cross-linguistic disparities by shaping encoding efficiency and surface inconsistency across writing systems.

Orthography concerns how linguistic content is represented in writing. Writing systems differ in *granularity*—whether symbols correspond roughly to phonemes, syllables, or morphemes—and in *transparency*, or how predictably written forms map to sounds (Wydell and Butterworth, 1999; Ziegler and Goswami, 2005). In humans, these differences can influence the difficulty of literacy acquisition, rather than spoken-language learning itself (Verhoeven and Perfetti, 2022; Chang et al., 2020). Although children learn to read at different rates across languages (Lai et al., 2024; Seymour et al., 2003), skilled adult reading is broadly similar across orthographies (Schroeder et al., 2022; Liversedge et al., 2016).

Language models, however, acquire language directly from text, without prior phonological, lexical, or semantic knowledge. Thus, orthographic differences matter mostly as differences in representation efficiency—how meaning is encoded into bytes and tokens. We focus on three consequences of orthography for modeling: encoding efficiency, vocabulary allocation in multilingual tokenizers, and the surface consistency of written forms.

First, writing systems differ in how much information they express per written symbol (Huang et al., 2024) and in how those symbols are encoded under UTF-8 (Yergeau, 2003). Alphabetic systems such as English distribute meaning across sequences of letters that roughly track phonological units, whereas logographic systems such as Chinese often express comparable content with fewer, denser characters (Tan et al., 2001). In abugidas such as Devanagari, characters are organized

around consonantal bases, with vowels often expressed through attached diacritics (Velayuthan and Sarveswaran, 2025). These differences lead to disparities at the byte level: English characters occupy one byte, Arabic characters two, and Chinese characters three; in Devanagari, what readers perceive as a single written unit may consist of a base consonant plus multiple diacritics, each separately encoded in three bytes (Lemire and Muła, 2022; Lavanya et al., 2005).

This creates a *byte premium*: under a fixed token or context budget, equal numbers of bytes do not correspond to equal amounts of linguistic content across languages (Arnett et al., 2024). For languages written in multi-byte scripts, the same content occupies longer encoded sequences, reducing effective exposure during training and shrinking the amount of text that fits within a context window (Moon et al., 2025). The problem is not only that some scripts yield longer sequences, but that equal budgets systematically allocate unequal amounts of usable input across languages.

Second, orthography affects how efficiently tokenizers can build reusable units. Most modern language models operate on subword vocabularies learned from corpus statistics, often via byte-pair encoding (BPE; Sennrich et al., 2016) starting from a byte-level vocabulary. Tokenization efficiency therefore depends not only on how much information each written symbol carries, but also on how easily recurring sequences can be merged into reusable units, which tends to disadvantage scripts whose characters are encoded with multiple bytes (Sennrich et al., 2016; Zouhar et al., 2023b; Kargaran et al., 2024). Thus, even under the same vocabulary budget, this can produce large disparities in compression across languages.

In multilingual settings, shared-vocabulary tokenization can allocate capacity unevenly: high-resource Latin-script languages tend to receive larger and more informative subwords, while many non-Latin scripts are segmented into shorter fragments with less information per token (Petrov et al., 2023; Ahia et al., 2023). Conversely, sharing or aligning scripts can improve transfer. For unseen or low-resource languages, transliteration into a script already well represented in the model can improve downstream performance (Muller et al., 2021; Moosa et al., 2023); it can also improve cross-lingual alignment, while recent work identifies script mismatch itself as a major barrier to cross-script knowledge transfer (Moosa et al., 2023; Bandarkar et al., 2026).

Byte-level tokenization also introduces distortions specific to multi-byte scripts. Because BPE merges frequent byte sequences rather than linguistically meaningful units, tokens may split characters across byte boundaries or contain only partial UTF-8 sequences (Firestone et al., 2025; Jang et al., 2025; Land and Arnett, 2025). Such fragments need not correspond to phonological, morphological, or semantic structure. Unrelated symbols may therefore partially overlap in tokenization simply because they share bytes, while meaningful sub-character structure may be obscured when token boundaries fail to align with it (Haslett, 2025). A straightforward character-level vocabulary would avoid some of these artifacts, but in multilingual settings even a base vocabulary of Unicode characters would already be extremely large—on the order of 130k types (Petrov et al., 2023; Ahia et al., 2023). Consequently, high information density per character does not necessarily yield equally efficient tokenization.

Third, orthography can shape the surface consistency of the training signal, because the same underlying content may appear in multiple written forms. This can arise from optional diacritics, as in Arabic and Hebrew (Inoue et al., 2026; Gorman and Pinter, 2025); from widespread spelling inconsistencies (Obeid et al., 2020; Adouane et al., 2019); and from routine script alternation, such as Simplified versus Traditional Chinese (Lyu et al., 2025) or South Asian languages commonly written in both native and Latin scripts (Roark et al., 2020). As a result, models may observe the same meaning dispersed across several orthographic variants rather than concentrated in a single stable form.

Tokenizer-free models can reduce some of these disparities (Pagnoni et al., 2025; Clark et al., 2022), and recent methods such as MYTE mitigate script-specific penalties by introducing alternatives to UTF-8 (Limisiewicz et al., 2024; Land and Arnett, 2025). However, these approaches do not eliminate orthographic asymmetries: byte premiums still lengthen sequences even for tokenizer-free models, character granularity remains incomparable across scripts, and surface variation can still fragment training signal. Overall, orthography contributes to cross-linguistic disparities by making encoding efficiency unequal across writing systems and written signal more or less consistent before modeling even begins.

2.2 Morphological Complexity

Apparent performance gaps from rich morphology often stem from segmentation quality, vocabulary budget, and data allocation.

Morphological complexity concerns how languages change or combine words to express distinctions such as tense, number, case, or word meaning, through processes such as inflection, derivation, and compounding (Haspelmath and Sims, 2013). Children ambiently learn the word-building patterns of their native language, although the pace of acquisition varies with the regularity and typology of the system (Clark, 2017). For adult second-language learners, these patterns can remain difficult, especially when they differ substantially from those of the learner's first language (Ellis, 2022). Perhaps because of this, morphology has often been assumed to increase language modeling difficulty (Cotterell et al., 2018; Gerz et al., 2018; Park et al., 2021; Mielke et al., 2019).

A common explanation is sparsity: morphologically rich languages realize each lexeme in many surface forms, lowering the frequency of individual forms and increasing the burden on models to generalize across paradigms even when the underlying rules are regular (Park et al., 2021). Early multilingual studies seemed to support this view: Cotterell et al. (2018) found that lower performance correlates with morphological richness across 21 languages, and that this effect was largely removed by lemmatization. Similarly, Gerz et al. (2018) reported substantial morphology-related differences across 50 typologically diverse languages.

Later work, however, suggests that much of this apparent effect is not intrinsic to morphology itself. Instead, it is amplified by how current pipelines segment, encode, and sample morphologically complex languages (Mielke et al., 2019). In particular, morphology-aware segmentation substantially reduces surprisal or performance gaps induced by standard BPE (Park et al., 2021; Mager et al., 2022), and analyses of WordPiece and BPE show that they often fail to preserve morpheme structure in languages with complex inflection or derivation (Klein and Tsarfaty, 2020; Lerner and Yvon, 2025). Recent work further shows that once tokenization or effective exposure are controlled, morphological complexity is a much weaker predictor of LM performance than previously assumed (Arnett and Bergen, 2025; Asgari et al., 2025; Rust et al., 2021).

Taken together, these results suggest that morphology affects language modeling through three interacting mechanisms. First, tokenization determines whether recurring morphemes are preserved as reusable units or broken into arbitrary fragments; when morpheme boundaries are obscured, models can generalize less effectively across related word forms (Mager et al., 2022; Park et al., 2021; Bostrom and Durrett, 2020; Gazit et al., 2025). Second, rich morphology can increase sequence cost when grammatical information is distributed across more fragmented token sequences, so the same content consumes more of the model's context, and equal token budgets result in less effective data exposure (Arnett et al., 2024; Foroutan et al., 2025; Asgari et al., 2025). Third, rich morphology can spread training signal across many low-frequency forms, while multilingual vocabulary learning may allocate less useful capacity to the morphemes needed to represent them, leaving less training signal for each individual form (Reif et al., 2025; Park et al., 2021; Rust et al., 2021).

Morphology is therefore best treated as an interaction effect rather than a uniform source of difficulty for language modeling. The same typological feature can appear harmful under one tokenizer or training budget and largely disappear under another. When these factors are removed (e.g., exposure is normalized for sequence-length and vocabulary-allocation effects) the gap between morphologically simpler and richer languages becomes substantially smaller.

2.3 Lexical Diversity and Vocabulary Size

Lexical diversity effects reflect tokenization misalignment, not linguistic complexity.

Lexical diversity captures how many distinct lexical types (lexemes and multiword expressions) a corpus contains and how evenly their frequencies are distributed. In human language acquisition, learning is strongly frequency-driven: high-frequency forms are acquired earlier, while low-frequency items are acquired later and remain harder to access (Ambridge et al., 2015), reflecting the long-tailed (Zipfian) distribution of words (Zipf, 1935). This connects lexical diversity to learnability: larger effective vocabularies entail longer tails of rare items, raising sample complexity for learning word meanings even if speakers ultimately master them.

Cross-linguistic differences in lexical diversity reflect lexicalization choices (what is expressed as a single word versus a multiword expression) and word-formation productivity (derivation and compounding) (Booij, 2005; Baayen, 2009). For instance, languages differ in how motion events are lexicalized (e.g., encoding manner versus path in the verb) (Talmy, 2000; Allen et al., 2007). Lexical diversity is typically measured from word-segmented corpora via type-frequency distributions, using indices like Type-Token Ratio and its length-normalized variants (Covington and McFall, 2010; McCarthy and Jarvis, 2010; Kettunen, 2014).[1]

In multilingual LM analyses, lexical diversity predicts perplexity and transfer quality (Mielke et al., 2019; Pelloni et al., 2022). However, perplexity alone does not reveal which linguistic attributes are learned (Meister and Cotterell, 2021). Output-side analyses also examine linguistic diversity in generations: Guo et al. (2025) evaluate model outputs along lexical, syntactic, and semantic diversity dimensions and find that current LLMs fall short of human-level linguistic diversity.

More generally, lexical diversity is a robust predictor of difficulty for LMs: Head-POS entropy (Dehouck and Denis, 2018) and raw type counts can outperform typological features in predicting language modeling difficulty (Mielke et al., 2019), and tokenization-sensitive measures such as Subword Evenness predict cross-lingual transfer and multilingual perplexity (Pelloni et al., 2022). Vocabulary-richness features also predict GPT-2 perplexity in English and interact with segmentation choices across typologies (Miaschi et al., 2021; Parra, 2024).

However, much of this effect reflects segmentation artifacts: in morphologically complex languages, frequency-based subwords fragment words into many pieces, inflating sequence length and reducing effective exposure per unit of semantic content (Lundin et al., 2025). When training data is equalized by byte premium or when tokenization artifacts are otherwise controlled, apparent lexical-diversity effects weaken substantially (Arnett and Bergen, 2025). Lexical diversity, therefore, challenges LMs mainly under tokenization schemes that misalign with linguistic structure. Vocabulary size remains a strong predictor, mainly due to segmentation and data sparsity rather than inherent lexical complexity.

[1] In corpus linguistics, these indices typically treat "tokens" as word tokens in a word-segmented corpus (not subword tokens produced by NLP tokenizers).

2.4 Syntactic Features

Syntactic features affect modeling difficulty indirectly through interactions with morphology, vocabulary size, and tokenization.

Syntactic features describe how languages organize words into phrases and clauses, including word order, case marking, and dependency structure. Syntax and morphology often provide alternative encodings for the same grammatical distinctions: a language may rely more on word order or more on overt marking (case, agreement) to signal roles and relations while preserving overall communicative efficiency (Sinnemäki, 2008; Lian et al., 2023; Levshina, 2021; Fedzechkina et al., 2017). In humans, these trade-offs influence which cues must be tracked rather than creating a global difficulty hierarchy.

The evidence on the effects of word order and syntactic variation on language modelling difficulty remains mixed (Mielke et al., 2019; Miaschi et al., 2021), and syntax is generally less investigated than morphology and tokenization in this context. Analyses based on typological features typically find that syntactic typology explains less variance in surprisal and perplexity than tokenization or lexical measures, with the largest effects occurring when critical syntactic cues rely on morphemes that subword tokenizers fragment (Mielke et al., 2019). Case marking illustrates this interaction: under standard BPE, languages with productive case systems show higher surprisal, but morphology-aware segmentation reduces the gap by segmenting case morphemes more consistently (Park et al., 2021), increasing their effective frequency and preserving cues for syntactic roles. Word order effects are mixed: basic order alone is not a reliable predictor of perplexity (Mielke et al., 2019), and reducing word-order-specific encoding can improve cross-lingual adaptation (Liu et al., 2021). Dependency distance metrics or embedding depth could, in principle, affect language modelling difficulty, but to the best of our knowledge, they have not yet been studied in this context.

In sum, syntactic differences rarely govern language modelling difficulty when taken in isolation; rather they interact with tokenization artifacts that inflate sequence length or obscure morphological

cues (Arnett and Bergen, 2025). Consequently, syntax-related performance gaps often reflect architectural constraints and English-centric positional heuristics rather than inherent modelling difficulty. Cognitively-motivated inductive biases, such as relative position encodings and syntactically informed attention, can mitigate these issues (Shaw et al., 2018; Dufter et al., 2022; Strubell et al., 2018; Kuribayashi et al., 2024), but positional design choices matter and multilingual evidence for newer schemes (ALiBi, RoPE) remains mixed (Ravishankar and Søgaard, 2021; Press et al., 2022; Su et al., 2024).

Overall, syntactic variation shapes cross-linguistic gaps mainly through interactions with morphology, tokenization, and vocabulary size; once normalized, syntax alone explains less of the variance, though it remains important for generalization and cross-lingual transfer.

2.5 Information-Theoretic Measures

Entropy differences largely capture representational choices, like word length or morphological encoding, rather than the intrinsic learnability of a language.

Information-theoretic metrics quantify predictability and redundancy, but they also reflect morphology, orthography, and other representational choices rather than pure learnability. Some potentially informative metrics remain difficult to define or measure, leaving room for future work. Information-theoretic measures quantify predictability and redundancy: *entropy* captures average uncertainty, *surprisal* measures the negative log probability of an observed unit, and *compression rate* approximates achievable code length under efficient encoding. These metrics provide a principled way to compare languages in terms of predictability and coding efficiency, linking cross-entropy in LMs to fundamental data statistics (Shannon, 1948). In human processing, surprisal theory formalizes the connection between predictability and cognitive difficulty (Hale, 2001; Smith and Levy, 2013).

A central insight from psycholinguistics and quantitative linguistics is that languages maintain stable information rates through compensatory trade-offs. Spoken languages converge on near-constant bits-per-second rates (Coupé et al., 2019; Jaeger, 2010), and morphologically rich languages exhibit higher per-word entropy because they encode more information per word (Bentz et al., 2017; Koplenig et al., 2025).

Large-scale studies show systematic differences in entropy at the character and word levels, balanced by structural features like word length (Koplenig et al., 2025). This reflects *Uniform Information Density* (UID), where languages spread information to keep local surprisal relatively stable (Jaeger, 2010; Levy and Jaeger, 2006), though UID might not be a universal law (Meister et al., 2021).

For LMs, entropy interacts with tokenization and sampling: high-entropy sequences require more data, and token-based budgets can exaggerate difficulty when scripts or tokenizers inflate sequence length. Byte-inefficient scripts and fragmented tokenization can inflate apparent entropy without adding semantic content (Rust et al., 2021). At the human-processing level, LM surprisal estimates can predict reading times across multiple languages, suggesting that surprisal is a useful–but imperfect–proxy for cognitive difficulty in cross-linguistic comparisons (Levy, 2008; Goodkind and Bicknell, 2018; Hollenstein et al., 2021; de Varda and Marelli, 2022; Wilcox et al., 2023; Kuperman et al., 2025).

Controlling for encoding efficiency and tokenization substantially reduces cross-linguistic surprisal gaps and narrows perplexity differences, indicating that part of the observed entropy variation reflects representation and sampling confounds (Arnett et al., 2024; Rust et al., 2021; Foroutan et al., 2025; Tsvetkov and Kipnis, 2024). However, perplexity remains an imperfect proxy for downstream performance: low perplexity can coexist with weak robustness, particularly in low-resource settings (Luitel et al., 2025; Gurgurov et al., 2025; Zhuang and Sun, 2025; Liu et al., 2022; Lourie et al., 2025).

Compression-based metrics provide architecture-independent baselines by evaluating predictability at fixed representational units. Bits-per-character/byte (BPC) estimates cross-entropy per character/byte, reducing reliance on subword tokenization and enabling comparisons that align with LM perplexity and transfer (De Souza et al., 2024; Tsvetkov and Kipnis, 2024). However, BPC is encoding-sensitive: UTF-8 byte premiums and script granularity distort comparisons even after byte normalization (Arnett et al., 2024; Moon et al., 2025; Foroutan et al., 2025; Deletang et al., 2024).

In sum, information-theoretic differences re-

flect language encoding choices rather than inherent learnability, and normalizing for density, byte length, or morphemes reduces many cross-linguistic gaps.

2.6 Typological Distance

> Typological diversity can cause difficulty in shared-parameter & -vocabulary settings.

Let's now move from single-language difficulty to cross-linguistic transfer, where patterns learned in one language can improve performance in others.

First, languages differ in many ways; this diversity represents alternative solutions to similar communicative constraints (Bickel, 2015; Comrie, 1989; Ponti et al., 2019). We can measure the difference between a pair of languages via their typological distance, which captures similarity in grammar (syntax, morphology), lexicon (cognates, word choice), and phonology, or via genealogical relatedness, which reflects shared ancestry.

In human L2 acquisition, linguistic distance predicts attainment and learning difficulty (Chiswick and Miller, 2005; Isphording and Otten, 2014; Schepens et al., 2020). Similar results have been suggested in models. Early multilingual models show that shared vocabularies bias representations toward related languages (Pires et al., 2019; Conneau et al., 2020b), with mBERT organizing languages along genealogical lines (Rama et al., 2020).

At a finer level, WALS-based similarity (Dryer and Haspelmath, 2013) predicts transfer quality beyond raw resource size (Lin et al., 2019), with features like word order and head direction particularly predictive (K et al., 2020; Blaschke et al., 2025). Tokenization-based diagnostics like Subword Evenness (Pelloni et al., 2022) and information-theoretic metrics like Information Parity (Tsvetkov and Kipnis, 2024) can predict cross-lingual transfer. Vocabulary overlap can sometimes predict positive transfer but sometimes be detrimental, depending on the exact task (Limisiewicz et al., 2023). For example Kallini et al. (2025) found that even vocabulary overlap of semantically unrelated words can be useful.

At larger scales, the *curse of multilinguality* refers to declining per-language performance as more languages share parameters (Conneau et al., 2020a), with low-resource and typologically distant languages suffering most (Lauscher et al., 2020). Typological distance amplifies interference and exacerbates vocabulary fragmentation. Controlled studies show that adding related languages improves low-resource performance, but may surprisingly hurt performance on high-resource languages (Chang et al., 2024a). Gradient conflicts are common when distant languages are trained jointly (Wang et al., 2020).

In summary, shared-parameter training with similar languages can help, but can also induce interference as typological diversity grows. Modular approaches that allocate language-specific capacity reduce conflict while preserving positive transfer (Pfeiffer et al., 2022; Blevins et al., 2024).

2.7 Summarizing Linguistic Properties

Across features, many performance gaps arise from mismatches between linguistic structure and modeling choices rather than intrinsic language difficulty. Tokenization and encoding can fragment cues and lengthen sequences, sampling can create unequal exposure, and shared-parameter training can cause negative transfer when typological diversity exceeds capacity. These factors also confound evaluation: low perplexity does not guarantee robust downstream performance, especially in low-resource settings. When segmentation, encoding, and exposure are normalized, many apparent cross-linguistic gaps shrink, showing that current modeling paradigms, not linguistic diversity itself, drive much of the disparity.

3 Design Implications

These findings motivate design implications for tokenization, sampling, architecture, evaluation, and corpus construction; we focus on interventions most directly supported by the surveyed evidence.

3.1 Tokenization: From Frequency-Based Segments to Linguistically Informed Units

Tokenization is one of the main design levers through which cross-linguistic disparities are either amplified or mitigated in multilingual language models. Frequency-based subword algorithms such as BPE and WordPiece often fragment morphemes and disproportionately disadvantage multi-byte scripts, inflating sequence length and compute cost while obscuring linguistically meaningful units (Park et al., 2021; Ali et al., 2024; Land and Arnett, 2025; Petrov et al., 2023; Arnett et al., 2024; Lundin et al., 2025). Across studies,

segmentation quality explains a substantial share of cross-linguistic performance differences, and morphology-, script-, and encoding-aware methods improve performance and efficiency across languages (Limisiewicz et al., 2024; Asgari et al., 2025; Mager et al., 2022).

Assessing tokenization quality remains nontrivial. Common diagnostics include compression, sequence length, corpus token count, vocabulary-balance measures, and related distributional metrics, but improvements on these measures do not always translate directly into better downstream performance (Schmidt et al., 2024; Zouhar et al., 2023a; Goldman et al., 2024; Dagan et al., 2024; Gallé, 2019). Tokenization should therefore be evaluated jointly in terms of efficiency, downstream utility, and parity across languages.

The surveyed work points to three broad intervention families. First, morphology-aware and language-adaptive tokenization can better preserve recurring morphemes and improve parity across languages (Asgari et al., 2025; Foroutan et al., 2025; Ahia et al., 2024). Second, alternative byte encodings can reduce disparities that arise when standard UTF-8 and shared-vocabulary BPE allocate capacity unevenly across scripts or create partial-byte artifacts (Limisiewicz et al., 2024; Land and Arnett, 2025). Third, tokenizer-free and character-level models can reduce script-specific tokenization penalties by avoiding fixed subword vocabularies altogether and reduce cross-lingual performance gaps, though typically at the cost of longer sequences and higher compute (Pagnoni et al., 2025; Clark et al., 2022).

Implication 1: Treat tokenization as a first-class multilingual design choice rather than a fixed preprocessing step. Prefer morphology-aware, script-aware, or language-adaptive tokenizers that better preserve meaningful units and allocate vocabulary capacity more evenly across languages. Consider alternative byte encodings to reduce systematic disadvantages introdcued by standard UTF-8 encoding. Tokenizer-free models can further reduce cross-lingual disparities, at the cost of longer sequences and higher compute.

3.2 Data Sampling and Byte Normalization

Token-based sampling penalizes byte-heavy scripts, whereas byte-normalized sampling narrows gaps (Arnett et al., 2024; Wei et al., 2021), motivating sampling strategies that target semantic exposure rather than raw token counts (e.g., UniMax, byte-premium scaling; Chung et al., 2023; Chang et al., 2024b; He et al., 2025).

Implication 2: Pretraining should use *byte-normalized*, *information-normalized*, or *morpheme-normalized* sampling for equal semantic coverage across languages. Data balancing should reflect linguistic diversity rather than corpus availability, correcting for segmentation bias, type proliferation, and script inefficiency.

3.3 Beyond One-Size-Fits-All Benchmarks

Current multilingual benchmarks often conflate linguistic difficulty with tokenization artifacts or dataset size. Perplexity is sensitive to tokenizer choice, whereas character-, morpheme-, and byte-level metrics provide more robust comparisons (Tsvetkov and Kipnis, 2024; Kanjirangat et al., 2025). Tokenizer-quality diagnostics and standardized reporting help disentangle measurement bias from true modeling capability (Chelombitko et al., 2024; Bender and Friedman, 2018).

Cross-linguistic syntactic challenge suites like CLAMS offer controlled tests of generalization and reveal consistent gaps between monolingual and multilingual models (Mueller et al., 2020). For morphology, community benchmarks (e.g. SIGMORPHON) provide fine-grained metrics that complement perplexity-based ones (Cotterell et al., 2017).

Probe-based evaluations show representational disparities, such as weaker subject/object identification in case-rich languages when models are trained on fixed word order (Papadimitriou et al., 2021), motivating typology-aware competency assessments.

Implication 3: Evaluation should use linguistically informed metrics and typology-aware probes beyond subword perplexity. Benchmarks should disaggregate performance by morphology, script, and word order to avoid masking inequities.

3.4 Balanced Corpora and Pretraining

Pretraining corpus choice strongly shapes multilingual performance. Web data overindexes English and underrepresents high-vitality languages, correlating poorly with global populations (Dunn, 2020; Dunn and Adams, 2020; Mehmood et al., 2017; Mor, 2025; Joshi et al., 2020; Khanna and Li, 2025; Bella et al., 2023). Corpus composition often tracks speaker counts over linguistic diversity, while data statements support accountable multilingual reporting (Bender and Friedman, 2018).

Multilingual corpora favor Indo-European languages and underrepresent complex morphology, minority scripts, and small populations. Balancing must account not only for token counts but also for linguistic density: information per token, morphological productivity, and rare-form distributions.

Resources such as UniMorph and high-coverage dependency treebanks can support typology-aware evaluation of coverage, even without direct training supervision (Nivre et al., 2020). These findings motivate moving beyond a single monolithic model: leveraging language similarity or tailoring components to typologically related clusters can boost learning for low-resource languages without forcing uniform representations (Malkin et al., 2022).

Implication 4: Corpus design should explicitly encode linguistic diversity by accounting for representational efficiency and linguistic density, ensuring that languages with high morphological or typological variation receive equivalent *semantic coverage*, not merely equivalent token counts.

4 Conclusions and Future Work

Multilingual performance is shaped less by inherent linguistic complexity than by design choices: tokenization, data allocation, and interference-aware training. Future work should explore language-adaptive strategies: predicting data and capacity needs per language, designing curricula that prioritize transfer from related languages, and developing architectures that dynamically allocate resources across typologically distinct languages. Truly low-resource and endangered languages require innovative approaches under scarcity.

Evaluation must also evolve: metrics should reflect cross-linguistic differences in task difficulty while capturing fairness and accessibility. Aligning model inductive biases with human learning can guide more robust multilingual NLP.

By embracing **linguistic diversity as a design principle**, we can build models that are more adaptable, equitable, and capable of supporting the full spectrum of the world's languages.

5 Limitations

Despite our analysis of multilingual performance, several limitations warrant consideration. First, our work focuses primarily on representation- and architecture-driven factors (tokenization, encoding, shared parameters) and does not fully capture other potential sources of difficulty, such as pragmatic, discourse-level, or sociolinguistic phenomena, which may affect real-world usage.

Second, most of our empirical insights rely on pretrained models and standard evaluation datasets, which may underrepresent truly low-resource or endangered languages. Data sparsity, orthographic variation, and non-standardized corpora in such languages could yield patterns not observed in higher-resource languages.

Third, while we consider cross-linguistic typology, our analysis is largely English-centric in architecture and benchmark design, which may bias conclusions about syntax, word order, and positional encoding effects.

Fourth, information-theoretic measures capture correlations with morphology and orthography rather than intrinsic learnability. Metrics for hierarchical structure, discourse-level predictability, or multimodal signals remain underexplored, leaving important aspects of language modeling outside our current framework.

Finally, despite our thorough literature survey, it is possible that relevant works were overlooked. We welcome pointers to such papers to keep this survey up to date.

Addressing these limitations in future work will be crucial for building truly language-adaptive, equitable, and robust multilingual models.

AI usage: The paper used AI assistance for rephrasing, for finding additional relevant papers, and occasionally for summarizing them.

References

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2019. Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 131–140, Hong Kong, China. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. *Advances in Neural Information Processing Systems*, 37:47790–47814.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *arXiv preprint arXiv:2305.13707*.

Doyin Akindotuni. 2025. Resource asymmetry in multilingual nlp: A comprehensive review and critique.

Journal of Computer and Communications, 13(7):14–47.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Shanley Allen, Aslı Özyürek, Sotaro Kita, Amanda Brown, Reyhan Furman, Tomoko Ishizuka, and Mihoko Fujii. 2007. Language-specific and universal influences in children's syntactic packaging of manner and path: A comparison of English, Japanese, and Turkish. *Cognition*, 102(1):16–48.

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Tyler A Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 1–9.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *arXiv preprint arXiv:2502.00894*.

R. Harald Baayen. 2009. Corpus linguistics in morphology: Morphological productivity. In *Corpus Linguistics: An International Handbook*, volume 2, pages 899–919. De Gruyter Mouton.

Lucas Bandarkar, Alan Ansell, and Trevor Cohn. 2026. Large reasoning models struggle to transfer parametric knowledge across scripts. *arXiv preprint arXiv:2603.17070*.

Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2023. Towards bridging the digital language divide. *CoRR*, abs/2307.13405.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, 19(6):275.

Ruth A. Berman. 2014. Cross-linguistic comparisons in child language research. *Journal of Child Language*, 41:26 – 37.

Balthasar Bickel. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10822–10837. Association for Computational Linguistics.

Geert Booij. 2005. Compounding and derivation: Evidence for construction morphology. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer, and Franz Rainer, editors, *Morphology and its Demarcations*, pages 109–132. John Benjamins Publishing Company.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin Bergen. 2024a. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024b. Goldfish: Monolingual language models for 350 languages. *CoRR*, abs/2408.10441.

Ya-Ning Chang, JSH Taylor, Kathleen Rastle, and Padraic Monaghan. 2020. The relationships between oral language and reading instruction: Evidence from a computational model of reading. *Cognitive Psychology*, 123:101336.

Iaroslav Chelombitko, Egor Safronov, and Aleksey Komissarov. 2024. Qtok: A comprehensive framework for evaluating multilingual tokenizer quality in large language models. *arXiv preprint arXiv:2410.12989*.

Barry R Chiswick and Paul W Miller. 2005. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of multilingual and multicultural development*, 26(1):1–11.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Eve V Clark. 2017. Morphology in language acquisition. *The handbook of morphology*, pages 374–389.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Bernard Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*, 2 edition. University of Chicago Press, Chicago, IL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL–SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. In *Forty-first International Conference on Machine Learning*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *ArXiv*, abs/2412.04261.

Leandro De Souza, Thales Almeida, Roberto Lotufo, and Rodrigo Frassetto Nogueira. 2024. Measuring cross-lingual transfer in bytes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7526–7537.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144. Association for Computational Linguistics.

Mathieu Dehouck and Pascal Denis. 2018. A framework for understanding the role of morphology in Universal Dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2864–2870, Brussels, Belgium. Association for Computational Linguistics.

Gregoire Deletang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. 2024. Language modeling is compression. In *The Twelfth International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Matthew S. Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2022. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Jonathan Dunn and Benjamin Adams. 2020. Mapping languages and demographics with georeferenced corpora. *arXiv preprint arXiv:2004.00809*.

Nick C Ellis. 2022. Second language learning of morphology. *Journal of the European Second Language Association*, 6(1).

Maryia Fedzechkina, Elissa L Newport, and T Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive science*, 41(2):416–446.

Preston Firestone, Shubham Ugare, Gagandeep Singh, and Sasa Misailovic. 2025. UTF-8 plumbing: Byte-level tokenizers unavoidably enable LLMs to generate ill-formed UTF-8. In *Second Conference on Language Modeling*.

Negar Foroutan, Clara Meister, Deepanway Paul, Joel Niklaus, and 1 others. 2025. Parity-aware byte-pair encoding: Improving cross-lingual fairness in tokenization. *arXiv preprint arXiv:2508.04796*.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Bar Gazit, Shaltiel Shmidman, Avi Shmidman, and Yuval Pinter. 2025. Splintering nonconcatenative languages for better tokenization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22405–22417, Vienna, Austria. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8920–8936, Suzhou, China. Association for Computational Linguistics.

Edward Gibson, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286, Bangkok, Thailand. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Kyle Gorman and Yuval Pinter. 2025. Don't touch my diacritics. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 285–291, Albuquerque, New Mexico. Association for Computational Linguistics.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages. *arXiv preprint arXiv:2502.10140*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

David A. Haslett. 2025. Tokenization changes meaning in large language models: Evidence from Chinese. *Computational Linguistics*, 51(3):785–814.

Martin Haspelmath and Andrea Sims. 2013. *Understanding morphology*. Routledge.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2025. Scaling laws for multilingual language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4257–4273, Vienna, Austria. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan

Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Linjieqiong Huang, Erik D. Reichle, and Xingshan Li. 2024. Comparative analyses of the information content of letters, characters, and inter-word spaces across writing systems. *Annals of the New York Academy of Sciences*, 1537(1):129–139.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nizar Habash, and Timothy Baldwin. 2026. Do diacritics matter? evaluating the impact of Arabic diacritics on tokenization and LLM benchmarks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 426–442, Rabat, Morocco. Association for Computational Linguistics.

Ingo E Isphording and Sebastian Otten. 2014. Linguistic barriers in the destination language acquisition of immigrants. *Journal of economic Behavior & organization*, 105:30–50.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61:23–62.

Eugene Jang, Kimin Lee, Jin-Woo Chung, Keuntae Park, and Seungwon Shin. 2025. Improbable bigrams expose vulnerabilities of incomplete tokens in byte-level tokenizers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18209–18216, Suzhou, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154.

Vani Kanjirangat, Tanja Samardzic, Ljiljana Dolamic, and Fabio Rinaldi. 2025. Tokenization and representation biases in multilingual models on dialectal NLP tasks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23992–24010, Suzhou, China. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotScript: A resource and tool for low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Saurabh Khanna and Xinxu Li. 2025. Invisible languages of the LLM universe. *CoRR*, abs/2510.11557.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209.

Alexander Koplenig, Sascha Wolfer, Jan Oliver Rüdiger, and Peter Meyer. 2025. Human languages trade off complexity against efficiency. *PLOS Complex Systems*, 2(2):1–42.

Victor Kuperman, Sascha Schroeder, Cengiz Acartürk, Niket Agrawal, Dominick Maia Alexandre, Lena Sophia Bolliger, Jan Brasser, César Campos-Rojas, Denis Drieghe, Dušica Filipović Đurđević, Luiz Vinicius Gadelha de Freitas, Sofya Goldina, Romualdo Ibáñez Orellana, Lena A. Jäger, Ómar I. Jóhannesson, Anurag Khare, Nik Kharlamov, Hanne B. S. Knudsen, Árni Kristjánsson, and 31 others. 2025. New data on text reading in english as a second language: The wave 2 expansion of the multilingual eye-movement corpus (meco). *Studies in Second Language Acquisition*, 47:677 – 695.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14522–14543.

Jialin Lai, Juan F Quinonez-Beltran, and R Malatesha Joshi. 2024. A bigger picture of early literacy and biliteracy acquisition in abugidas: Perspectives from asian and african languages. *Reading Research Quarterly*, 59(3):499–513.

Sam Land and Catherine Arnett. 2025. Bpe stays on script: Structured encoding for robust multilingual pretokenization. *arXiv preprint arXiv:2505.24689*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Prahallad Lavanya, Prahallad Kishore, and Ganapa Thiraju Madhavi. 2005. A simple approach for building transliteration editors for indian languages. *Journal of Zhejiang University-SCIENCE A*, 6(11):1354–1361.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Lemire and Wojciech Muła. 2022. Transcoding billions of unicode characters per second with simd instructions. *Software: Practice and Experience*, 52(2):484–508.

Paul Lerner and François Yvon. 2025. Unlike "likely","unlike" is unlikely: Bpe-based segmentation hurts morphological derivations in llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5181–5190.

Natalia Levshina. 2021. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology*, 12:648200.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy and T. Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Tomasz Limisiewicz, Jiri Balhar, and David Marecek. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of ACL 2023*.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. Myte: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2022. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, pages 13461–13469. AAAI Press.

Simon P Liversedge, Denis Drieghe, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä. 2016. Universality in eye movements and reading: A trilingual investigation. *Cognition*, 147:1–20.

Nicholas Lourie, Michael Y. Hu, and Kyunghyun Cho. 2025. Scaling laws are unreliable for downstream tasks: A reality check. ArXiv.

Nishan Luitel, Nishant Bekoju, Anil Kumar Sah, and 1 others. 2025. Can perplexity predict finetuning performance? an investigation of tokenization effects on sequential language models for nepali. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning*.

Jessica M. Lundin, Ada Zhang, Nihal Karim, Hamza Louzan, Victor Wei, David Adelani, and Cody Carroll. 2025. The token tax: Systematic bias in multilingual tokenization. *arXiv preprint arXiv:2509.05486*.

Hanjia Lyu, Jiebo Luo, Jian Kang, and Allison Koenecke. 2025. Characterizing bias: Benchmarking large language models in simplified versus traditional

chinese. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2815–2846, New York, NY, USA. Association for Computing Machinery.

Manuel Mager, Arturo Oncevay, Elisabeth Maier, Katharina von der Wense, and Thang Vu. 2022. Bpe vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Muhammad Asif Mehmood, Hafiz Muhammad Shafiq, and 1 others. 2017. Understanding regional context of world wide web using common crawl corpus. In *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, pages 1–6. IEEE.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021. What makes my model perplexed? a linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989.

Seonmin Moon, Tatsuya Hiraoka, and Naoaki Okazaki. 2025. Bit-level bpe: Below the byte boundary. *arXiv preprint arXiv:2506.07541*.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Niva Mor. 2025. It's a global village (if you speak the right language): On language models, digital sidelining, and participation. *Wisconsin International Law Journal*, 42:329.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srini Iyer. 2025. Byte latent transformer: Patches scale better than tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9238–9258, Vienna, Austria. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Isidro Parra. 2024. Morphological typology in bpe subword productivity and language modeling. *arXiv preprint arXiv:2410.23656*.

Olga Pelloni, Anastassia Shaitarova, and Tanja Samardzic. 2022. Subword evenness (sue) as a predictor of cross-lingual transfer to low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7428–7445. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vinit Ravishankar and Anders Søgaard. 2021. The impact of positional encodings on multilingual compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 763–777, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuval Reif, Guy Kaplan, and Roy Schwartz. 2025. Vocab diet: Reshaping the vocabulary of llms with vector arithmetic. *arXiv preprint arXiv:2510.17001*.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Job Schepens, Roeland Van Hout, and T Florian Jaeger. 2020. Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, 194:104056.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Sascha Schroeder, Tuomo Häikiö, Ascensión Pagán, Jonathan H Dickins, Jukka Hyönä, and Simon P Liversedge. 2022. Eye movements of children and adults reading in three different orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(10):1518.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL 2016*, pages 1715–1725.

Philip HK Seymour, Mikko Aro, Jane M Erskine, and Collaboration with COST Action A8 Network. 2003. Foundation literacy acquisition in european orthographies. *British Journal of psychology*, 94(2):143–174.

Claude E. Shannon. 1948. *A Mathematical Theory of Communication*. Bell System Technical Journal.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468.

Kaius Sinnemäki. 2008. Complexity trade-offs in core argument marking. *Language complexity*, pages 67–88.

Dan Slobin. 1987. The crosslinguistic study of language acquisition. *The Modern Language Journal*, 71:371.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Leonard Talmy. 2000. *Toward a Cognitive Semantics, Volume 2: Typology and Process in Concept Structuring*. MIT Press, Cambridge, MA.

Li Hai Tan, Ho-Ling Liu, Charles A Perfetti, John A Spinks, Peter T Fox, and Jia-Hong Gao. 2001. The neural system underlying chinese logograph reading. *Neuroimage*, 13(5):836–846.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989, Miami, Florida, USA. Association for Computational Linguistics.

Menan Velayuthan and Kengatharaiyer Sarveswaran. 2025. Egalitarian language representation in language models: It all begins with tokenizers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5987–5996, Abu Dhabi, UAE. Association for Computational Linguistics.

Ludo Verhoeven and Charles Perfetti. 2022. Universals in learning to read across languages and writing systems. *Scientific Studies of Reading*, 26(2):150–164.

Chenglong Wang, Haoyu Tang, Xiyuan Yang, Yueqi Xie, Jina Suh, Sunayana Sitaram, Junming Huang, Yu Xie, Pengjun Zhao, Zhaoya Gong, and 1 others. 2025. Uncovering inequalities in new knowledge learning by large language models across different languages. *Proceedings of the National Academy of Sciences*, 122(51):e2514626122.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Junqiu Wei, Qun Liu, Yinpeng Guo, and Xin Jiang. 2021. Training multilingual pre-trained language model with byte-level subwords. *arXiv preprint arXiv:2101.09469*.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Taeko Nakayama Wydell and Brian Butterworth. 1999. A case study of an english-japanese bilingual with monolingual dyslexia. *Cognition*, 70(3):273–305.

François Yergeau. 2003. UTF-8, a transformation format of ISO 10646. RFC 3629.

Raoyuan Zhao, Yihong Liu, Hinrich Schütze, and Michael A Hedderich. 2025. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. *arXiv preprint arXiv:2510.09555*.

Wei Zhuang and Yan Sun. 2025. Cute: A multilingual dataset for enhancing cross-lingual knowledge transfer in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Johannes C. Ziegler and Usha Goswami. 2005. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, 131 1:3–29.

George K. Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023a. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023b. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614, Toronto, Canada. Association for Computational Linguistics.